An Inventory and Procurement Policy for the Deep Space Network

I. Eisenberger, F. R. Maiocco, and G. Lorden[1]
Communications Systems Research Section

A technical description of a proposed inventory and procurement policy for ordering and allocating maintenance and operating supplies throughout the Deep Space Network is presented. This policy differs from the conventional economic lot-size procurement policy in that the reorder point for the Network Supply Depot (NSD) depends upon the stockage levels at all area station or Complex Supply Facilities (CSF), as well as on the level at the NSD. Thus, by basing reorder decisions upon the state of the inventory supplies throughout the entire DSN, an efficient cost minimizing policy is possible. Safe minimum inventory levels are established for each CSF by means of statistical decision theory techniques which require NSD to reorder whenever one or more of the CSFs reaches its prescribed minimum. Some results of a statistical study of the effect of this policy are included.

I. Introduction and Summary

The key feature of the following proposed inventory and procurement policy for the Deep Space Network is the continuous updating of data files in the Integrated Logistics System's (ILS) Supply Inventory Subsystem (SIS). These data files contain the following for each inventory item:

(1) The stockage level on hand at each CSF and at NSD.

(2) The (estimated) demand distribution at each CSF.

(3) Cost parameters associated with the item.

This updating is to be accomplished by utilizing the existing Complex and Network teletype facilities which access a central computer for updating data files via a remote terminal. Another means to update data files is to automate the transaction at each CSF at the time of issuance. Such data could be stored in a cassette tape and in a specified format for later access by the central computer. The transaction information would be available for file maintenance on an as-required or near-real-time basis.

Each time a demand for a particular item occurs at one or more of the CSFs (and their computer files are updated at NSD), the sufficiency of the current stockage level at a CSF is compared to a "safe" minimum stockage level for the particular CSF. This level is a function of the probability of being out of stock during one lead period and of

[1] Consultant, California Institute of Technology, Mathematics Department.

the cost parameters. If the current stock level at even one CSF falls below the minimum, a computer program determines

(1) The numbers of units to be shipped from the NSD to certain CSFs.

(2) If NSD is out of stock, the number of units to be ordered by the NSD from outside sources.

When a shipment is received at NSD, a program is run which prints out

(3) The number of units to be shipped to each CSF.

(4) The number of units (if any) to be placed in NSD inventory.

An alternative mode of operation which is slightly less effective is to make the computer run yielding (3) and (4) at the time the order is placed. The stock allocated to NSD in (4) is called the *reserve* stock. Two types of policies are considered below:

(1) No-reserve policy (NR-policy) under which all units are allocated and shipped to CSFs as soon as they are received at NSD.

(2) One-stage reserve policy (R-policy) under which a number of units determined in (4) are placed in NSD inventory, and all are shipped to CSFs as soon as one of the CSFs reaches its minimum level.

These policies are discussed in Section II. The choice between them depends on a tradeoff between the reduced CSF inventory levels (and, hence, costs) brought about by R-policies and the increased costs of handling and shipping from NSD to CSFs. It is anticipated that the no-reserve policy is preferable for many low-cost items.

Both policies are implemented through the use of an optimal allocation table (see Section II) which is based upon estimated mean demand rates at the CSFs. These estimates are simply the means of the demand distributions, which are continuously updated. Tables listing these optimal allocations are easily constructed for a variety of demand rates taking advantage of the fact that the allocations depend only on the ratios of CSF mean demand rates. For critical or costly items, the computational algorithm used to construct the tables can be used directly each time an allocation is necessary. Both the NR- and R-policies have another important feature in common: a minimum inventory level (greater than zero) is not maintained at NSD. Maintaining such a minimum incurs excessive inventory costs because the minimum may be reached and NSD reorders may be placed at a time when the levels of inventory on hand at the CSFs are large. An efficient cost-minimizing policy must base reorder decisions upon the state of the entire system, i.e., the levels of inventory on hand at the CSFs as well as the NSD level. It should be noted that the lead period used to determine safe CSF minimum levels includes the time for NSD to receive the order plus the shipping time to the CSFs.

The size of the NSD reserve stock (in the R-policy) tends to be small compared to the total quantity allocated to the CSFs, typically 10–15%. This occurs because the effectiveness of the R-policy in reducing average inventory costs depends upon all CSFs reaching (or approaching) their minimum levels at approximately the same time. The most effective size for the reserve stock tends to be in the range where it serves as a small "correction" to the chance fluctuations around the mean demand rates.

The method used to determine the size of an NSD order, also discussed in Section II, is a modification of the conventional economic lot size model (Refs. 1, 2) designed to take into account the effects of ordering simultaneously for all CSFs.

The question of when to order, i.e., determining the safe minimum levels for the CSFs, is taken up in Section III. The cost parameters determine what is called (in the following discussion) a *cost-criticality quantile.* An inventory level is considered safe if it exceeds the cost criticality quantile of the (current) demand distribution. (In Section III, the current estimate of the demand distribution is called the *posterior distribution* of Y, where Y is the demand during one lead period). Thus, the heart of the method of determining CSF minimum levels is the derivation of estimated demand distributions for the CSFs. These estimates are based on:

(1) Prior information regarding demand patterns for an item (e.g., "engineering judgment").

(2) The continuously updated demand experience at each CSF.

Source (1) may not be available for many items; in any case, its influence diminishes steadily as demand experience is accumulated.

A natural way of obtaining an estimate of the demand distribution is to assume that its mean is the observed mean demand and then use the standard model of a Poisson distribution with that mean. The inadequacy of this approach results not from the Poisson model (a generalized version of which is used in Section III), but rather from the unequal consequences of overestimating and underestimating the mean demand.

To illustrate, suppose the mean demand is actually 10 and the distribution of demand is given by the solid curve in Fig. 1.

The left-hand dotted curve represents the estimated demand distribution if (through chance fluctuation) one observes a mean demand of 7. The right-hand dotted curve comes from an observed mean demand of 13. For typical values of cost parameters, the cost-criticality quantile is around 0.90, so that shortages occur in only one-tenth of the lead periods. Thus if the left-hand dotted curve were assumed true, the minimum level s would be set at about 10, say, so that the shaded area is only 10% of the area under the curve. Recall, however, that the solid curve is the true demand distribution, so that the true probability of a shortage during any lead period would be about one-half. If an error of compatible magnitude were made in the opposite direction (i.e., overestimating the mean demand) one would have an estimated demand curve like the one on the right-hand side of Fig. 1 and would set s at about 16. As a consequence of this overestimation, an extra three units would be maintained in inventory, resulting in an added inventory cost. This added cost would have less impact than the very high shortage probability resulting from underestimated mean demand.

It is clear, therefore, that minimum levels should be estimated conservatively. To what degree and according to what formulas these levels should be estimated are discussed in Section III through the use of statistical decision theory. For the purpose of setting CSF minimum levels, all of the information derived from observed demand (and possibly prior information) can be summarized in the estimated (or posterior) demand distribution, which is easily updated each time a CSF demand is experienced.

Also included in Section III are some results of an extensive investigation which was carried out to evaluate the performance of the procedures developed for setting CSF minimum levels (i.e., deciding when to reorder). An additional aim of this investigation was to develop guidelines for choosing the three parameter values needed to specify such a procedure. The art of choosing these parameter values is one of the main topics of Section III. In application to the DSN, large categories of items would be subject to the same parameter choices.

II. Procurement and Allocation Policies

A. Inventory Concepts

The performance of any inventory policy is measured by the average cost it incurs per unit time and the frequency and severity of shortages. A convenient set of parameters to use in expressing the average cost is the following (all referring to a single inventory item):

M = mean demand per year for all CSFs combined

h = cost of stocking one unit for one year

K = fixed cost of preparing, handling, and shipping an order (excluding "per unit" costs)

T = mean time between reorders (in years)

W_i = average inventory at ith CSF when new shipment (not reserve shipment) is received.

Note that the last two parameters depend on the policy chosen while the first three do not. We are also concerned with the parameter

U_i = average shortage at ith CSF when reorder arrives. (When there is no shortage, "0" is averaged in.)

The average inventory cost can be expressed in terms of the specified parameters as

$$C = \frac{1}{2} MTh + \frac{K}{T} + h \sum_i W_i \qquad (1)$$

and this cost is incurred along with an average shortage

$$\sum_i U_i$$

Obviously, lower average shortage can be achieved by maintaining higher stockage levels, thus resulting in greater average cost, Eq. (1). The tradeoff between the two must be made on the basis of an assessment of the importance of shortages (see Appendix A on the cost criticality quantile). To help in making this assessment, one can use the methods of Appendix A to examine the relationship between average cost and expected shortage, i.e., to see how much reduction in expected shortage is "bought" by successive increments of average cost. In this Section and Section III, we are interested in developing a range of policies, all of which are efficient in the sense that their average cost cannot be improved without incurring greater expected shortage. A mathematical device for characterizing such policies is to introduce a criticality parameter in the form of a "cost":

p = cost per unit of shortage

and consider the overall average cost per year as

$$\frac{1}{2}MTh + \frac{K}{T} + h\sum_i W_i + \frac{p}{T}\sum_i U_i \qquad (2)$$

The efficient procedures, then, are those which minimize Eq. (2) for some value of $p > 0$. By partial differentation with respect to T, it is easily seen that

$$T = \left[\frac{2(K + p\sum_i U_i)}{Mh}\right]^{1/2} \qquad (3)$$

is necessary to minimize Eq. (2). The value of

$$\sum_i U_i$$

depends on the choice of minimum levels, which is considered in Section III. Since optimality in this choice depends on T, a standard iterative procedure is useful for determining both simultaneously (Ref. 2). For the ranges of parameter values typical of the DSN, the first term in the numerator in Eq. (3) dominates, so that

$$T = \left(\frac{2K}{Mh}\right)^{1/2} \qquad (4)$$

is a satisfactory approximation.

Given the value of T desired, we have the problem of choosing the quantity to be ordered to achieve that value of the expected time until reorder is necessary. Assume for the time being that the following condition is imposed: shipments arriving at NSD are immediately allocated and shipped to the CSFs, with no inventory maintained at NSD. (As discussed below, this approach is a reasonable one for many low-cost items.)

B. Allocation Problem for NR-Policies

Let $t_1^i, t_2^i, t_3^i, \cdots$ be the (random) waiting times between successive demands at the ith CSF. The t_j^i's are assumed independent and exponentially distributed with mean $1/\lambda_i$ depending on the CSF (estimated from demand experience). It is assumed that each demand is for a single item. (This simplifying assumption is relaxed for the considerations of Section III, but to do so here would unnecessarily complicate the computations.) Given $n_1, \cdots, n_k$ items allocated to the CSFs (in excess of their minimum stockage levels), the expected time until one of the CSFs reaches its minimum is

$$E\min_i (t_1^i + \cdots + t_{n_i}^i) = \int_0^\infty \prod_{i=1}^k F_{\lambda_i t}(n_i - 1)\,dt \qquad (5)$$

where $F_{\lambda t}$ denotes the Poisson distribution function with mean λt. It is not necessary to compute Eq. (5) for all possible k-tuples $(n_1, \cdots, n_k)$—only for k-tuples which are efficient in the sense that they maximize

$$E\min_i.(t_1^i + \cdots + t_{n_i}^i)$$

among all k-tuples with the same initial inventory total, $n_1 + \cdots + n_k$. (By a Martingale systems theorem, this efficiency criterion can be shown equivalent to minimizing the expected total surplus inventory, $\sum_i W_i$.)

Let Q denote $n_1 + \cdots + n_k$. A simple computer program evaluates Eq. (5) numerically and optimizes the choice of $n_1, \cdots, n_k$ recursively for $Q = k, k+1, \cdots$. Given the optimal allocation for $Q = \ell$, the optimal allocation for $Q = \ell + 1$ is obtained by computing Eq. (5) for each allocation giving one more unit to some CSF than the allocation for $Q = \ell$. There are k such allocations and the one yielding the largest value in Eq. (5) is selected. Thus one obtains an allocation table of the Table 1 type (easily stored in computer memory).

The "Expected time" column denotes the quantity in Eq. (5). The value of Q is chosen so that

$$\text{Expected time} \approx \text{optimal } T \text{ (mean time between orders)} \qquad (6)$$

as closely as possible. When a reorder is necessary (i.e., at least one CSF is at its minimum) the NR-policy calls for an order size

$$\text{Order size} = Q - \sum_i \text{(residual at } i\text{th CSF)} \qquad (7)$$

where the *residual* at a CSF is the inventory level at reorder time minus the minimum inventory level (e.g., zero, for the CSF that reaches minimum). The "Expected residual" column in the allocation table gives the average value of the total of these residuals, i.e., the average value of the quantity subtracted from Q in Eq. (7) to obtain the order size. Equation (7) allows for each CSF to receive as its share of the quantity ordered:

$$n_i - \text{residual at } i\text{th CSF} \qquad (8)$$

where n_i is the portion of Q specified in the table. Increased efficiency of allocation can be achieved by recomputing the allocation when the new shipment arrives, thus taking into account the actual CSF inventory levels at that time. *Example:* Suppose the residuals at the CSFs at reorder time are 4,0,7,5,8,2, respectively, and $T = 1.50$ is desired. For $Q = 117$, Table 1 gives expected time 1.50 and allocates 16,17,19,20,22,23. Thus $117 - 26 = 91$ items are ordered and are to be allocated to the CSFs by Eq. (8) as follows: 12,17,12,15,14,21.

If recomputation is made when the order arrives at NSD, then the allocation can take into account the actual inventory levels at that time. Suppose these current inventory levels (with each CSF minimum subtracted) are 2, − 3,4,5,6,0, reflecting demands during lead period 2,3,3,0,2,2, respectively. Now the actual quantity available for allocation is $91 + (2 - 3 + 4 + 5 + 6) = 105$, and Table 1 indicates the allocation of 14,15,17,18,20,21. Subtracting the current inventory levels (in excess of minimum), we find that the 91 items are to be allocated 12,18,13,13,14,21, which differs slightly from the results in the preceding paragraph.

C. Improved Allocation With R-Policies

We now consider some possible improvements in allocating supplies to the CSFs when the no-reserve condition is dropped and an NSD inventory can be utilized. These improvements come about by reduction of the surplus inventory term

$$h \sum_i W_i$$

in the average cost equation, Eq. (2). However, any resupply incurs additional shipping and handling (but not ordering) costs, including the cost of placing items in NSD inventory and taking them out. These additional costs must be assessed and weighed against the expected savings in the surplus inventory term.

For low-cost items stocked in small or moderate quantities, it is clearly disadvantageous to handle an NSD supply merely to effect a small savings in CSF surplus inventory cost. For more expensive or higher-volume items, the potential savings are worth analyzing and hence a specific alternative to the no-reserve policy must be formulated.

The simplest alternative, which would incur the least increase in shipping and handling costs, is a one-stage reserve policy (R-policy) which operates as follows. A quantity R from each incoming order is placed in NSD inventory, the remainder being allocated and shipped to the CSFs. As soon as one CSF reaches its minimum stockage level, the entire reserve stock R is allocated and shipped to the CSFs (some of them possibly receiving zero). The next time one of the CSFs reaches minimum, NSD must reorder.

The time required to ship from NSD to the CSFs is on the order of one-half the total of NSD lead period for reorder plus shipping to CSFs. Hence, the expected shortage during the waiting period for the R-supply is usually negligible. The average surplus inventory W_i is approximately equal to the average residual inventory at the ith CSF plus the minimum level, minus the mean demand during lead time. This occurs because in a typical lead period there is no shortage. Consequently, the reduction in

$$\sum_i W_i$$

resulting from an R-policy can be approximated by reduction in the total expected residual inventory (excess over minimum levels) at reorder time.

Extensive investigations of R-policies were carried out by Monte Carlo simulation on the XDS Sigma V computer and the following rules were found to be most efficient:

(1) Choose Q and the order size as in the NR-policy (i.e., Eqs. 6 and 7).

(2) Choose $R = 1/2$ expected residual in the allocation table, Table 1.

(3) After subtracting R from the total quantity available for allocation, allocate the rest using Table 1 as in the NR-policy.

One more procedure needs to be specified: how to allocate the supply R among the CSFs. At the time the reserve stock is to be shipped, NSD has R units, at least one CSF has zero units in excess of its minimum level, and the other CSFs have relatively small numbers of units. In this situation the available total quantity, R plus the total excess over CSF minimum levels, often cannot be allocated according to Table 1. For example, if $R = 17$ and the CSF levels (over minimum) are 3,7,1,4,0,8, respectively, then the total quantity available is 40, and Table 1 calls for the CSFs to have 6,6,6,7,7,8 (over minimum), respectively.

But complex number 2 already has more than called for, and since (we are assuming) CSFs do not ship supplies to other CSFs, an alternative allocation must be made. The optimum alternative allocation subject to the constraint that every CSF is allocated no less than its current level can be calculated by an algorithm similar to the one used for the allocation table.

However, almost exactly the same results can be achieved much more simply by a modified use of the allocation table. The current quantities on hand at the CSFs are entered in the table in place of all smaller entries. Thus none of the allocations specify fewer units for a CSF than it already has. The allocation to be used is the one in the table for which the total of the CSF allocations agrees with the number of units actually available.

In the example above, row 40 in Table 1 would be changed to 6,7,6,7,7,8, for a total of 41. Row 39 in Table 1 (not shown) would be changed from 5,6,6,7,7,8 to 5,7,6,7,7,8. The latter allocation adds up to 40 and hence would be the one used.

D. Example of Comparison Between NR-Policy and R-Policy

Suppose the row of Table 1 with $Q = 142$ is used for the NR-policy. Then, according to the table, the expected residual at reorder time is 34.8. Using the R-policy with $R = 17$, the expected residual averages 19.9. The net reduction of 14.9 is 10.5% of the initial quantity, 142, and the average inventory cost, Eq. (1), is reduced by the same percentage.

III. CSF Minimum Stockage Policies

A. General

The problem of setting minimum levels for a CSF inventory consists of minimizing the part of the average cost equation (Eq. 2) that involves the minimum levels. That part is

$$h\sum_i W_i + \frac{p}{T}\sum_i U_i = \sum_i \left(hW_i + \frac{p}{T}U_i\right) \qquad (9)$$

where h, p, and T are known (see Section II) and

W_i = average inventory level at ith CSF when reorder arrives

U_i = average shortage (unfilled demands) at ith CSF when reorder arrives.

The sum over i in Eq. (9) is minimized by choosing separately for each CSF the minimum stockage level that minimizes

$$hW + \frac{p}{T}U \qquad (10)$$

The subscripts in Eq. (10) are dropped since we will consider from now on a fixed CSF.

To find the optimum value of

s = minimum stockage level

it is necessary to express W and U in terms of the random variable

Y = total demand (in number of items) at the CSF during one lead period.

The number of units on hand at the CSF at the beginning of a lead period may be greater than s (if another CSF reached minimum first) and may be less than s (if an order for more than one unit reduces the inventory level below minimum). To obtain approximate expressions for W and U it will be assumed that the inventory level at the beginning of a lead period equals s. Then

$$W = E(s - Y)^+$$

and

$$U = E(Y - s)^+$$

where E denotes the expected value or average and "+" means "positive values of," the negative values being averaged in as zero. (If, for example, $Y - s$ is negative, then there is no shortage and zero is averaged into the determination of U.) Therefore, the problem is to minimize

$$hE(s - Y)^+ + \frac{p}{T}E(Y - s)^+ \qquad (11)$$

which equals

$$hE(s - Y) + \left(\frac{p}{T} + h\right)E(Y - s)^+ \qquad (12)$$

In order to obtain a preliminary result, it will be assumed for the moment that the probability distribution of Y is known. To find the minimizing s in this case, let $B(s)$ denote the expression to be minimized (Eq. 12) and note that

$$\begin{aligned} B(s+1) - B(s) &= h + \left(\frac{p}{T} + h\right)\left\{\sum_{y=s+1}^{\infty}(y - (s+1))P(Y = y) - \sum_{y=s}^{\infty}(y - s)P(Y = y)\right\} \\ &= h + \left(\frac{p}{T} + h\right)\left(-\sum_{y=s+1}^{\infty}P(Y = y)\right) \\ &= \left(\frac{p}{T} + h\right)P(Y \leq s) - \frac{p}{T} \end{aligned}$$

Since $P(Y \leqq s)$ increases with s, so does $B(s+1) - B(s)$, and the minimizing s is obviously the smallest s such that $B(s+1) - B(s)$ is positive, or, what is equivalent,

$$P(Y \leqq s) > \frac{\dfrac{p}{T}}{\dfrac{p}{T} + h} \tag{13}$$

Thus the value of s minimizing $B(s)$ is the smallest s satisfying Eq. (13). Letting

$$q = \frac{\dfrac{p}{T}}{\dfrac{p}{T} + h} = \frac{1}{1 + \dfrac{hT}{p}} \tag{14}$$

the minimizing s is the qth quantile of the distribution of Y, the demand during lead time. This q is called the *cost-criticality quantile.*

The situation in the DSN is that the demand distribution for a CSF is unknown. However, its qth quantile can be estimated from observed demand over a time period t (measured in units of one lead period). We proceed, next, to describe a method for estimating this cost criticality quantile and thus the minimum level from observed data.

B. Estimation of the Cost Criticality Quantile q

We assume that demand is stationary over time and hence describable by a compound Poisson process. That is, orders of size $k\,(k = 1, 2, \cdots)$ arrive independently of each other with mean frequency λ_k per unit time and with probabilities

$$P\,(m \text{ orders of size } k \text{ during time } t) = \exp(-\lambda_k t)\,\frac{(\lambda_k t)^m}{m!} \tag{15}$$

$$\text{for } t > 0,\ m = 0, 1, \cdots$$

Let X_k = the number of orders of size k observed during time $t > 0$. Then from Eq. (15) we have

$$P(X_k = x) = \exp(-\lambda_k t)\,\frac{(\lambda_k t)^x}{x!}, \text{ for } x = 0, 1, 2, \cdots.$$

Let $\mathbf{X} = (X_1, X_2, \cdots)$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots)$. Assume the minimum stockage level s is to be chosen as a function of $\mathbf{X}$. Then the resulting average cost is

$$C(\boldsymbol{\lambda}) = hE_{\boldsymbol{\lambda}}(s(\mathbf{X}) - Y) + \left(\frac{p}{T} + h\right) E_{\boldsymbol{\lambda}}(Y - s(\mathbf{X}))^{+}$$

where the symbol $E_{\boldsymbol{\lambda}}$ is used to indicate that the expected values depend upon the true values of the λ_k's (e.g., for large λ_k's, both $\mathbf{X}$ and Y will be large with high probability).

It is desired to minimize $C(\boldsymbol{\lambda})$. But it is easily seen that no rule for choosing s can minimize $C(\boldsymbol{\lambda})$ for all $\boldsymbol{\lambda}$ simultaneously. (For different $\boldsymbol{\lambda}$s, the distribution of λ is different, and its qth quantile differs.) We can only choose $s(\mathbf{X})$ to be a good estimate of the qth quantile of the (unknown) distribution of Y, based on the observed $\mathbf{X}$ over time t. For each $\boldsymbol{\lambda}$, define

$$R(\lambda) =$$

$$C(\boldsymbol{\lambda}) - \min_{s \geqq 0} hE\left\{(s - Y) + \left(\frac{p}{T} + h\right) E_{\boldsymbol{\lambda}}(Y - s)^{+}\right\}$$

the *regret* (in the form of extra average cost) resulting from use of the estimate $s(\mathbf{X})$ rather than the (unknown) best possible s for $\boldsymbol{\lambda}$. A reasonable goal may be roughly defined as follows: for a broad range of $\boldsymbol{\lambda}$s, keep the regret small. This estimation problem, with the regret function (of $\boldsymbol{\lambda}$) given above, is an example of a statistical decision problem, in the sense of Wald. The general theory of such problems (Ref. 3) leads to the conclusion that essentially all worthwhile estimation rules are solutions for some probability density, $g(\lambda)$, of the following problem:

$$\text{Choose } s(\mathbf{X}) \text{ so as to minimize} \int R(\boldsymbol{\lambda})\, g(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}$$

In other words, the worthwhile estimation rules are those which minimize some *average* of the values of the regret function (the averaging being done according to a prescribed g). Such a minimizing rule is called a *Bayes solution* with respect to the density g, which is called an *a priori density.* The Bayes solutions of the present problem can be found for a large class of g's and the choice of an estimation rule among these solutions is easy to analyze. To do this, it is helpful to consider first the special case where all orders are of size one, so that $\mathbf{X}$ and $\boldsymbol{\lambda}$ are one-dimensional: X and λ. We have

$$P_\lambda(X = x) = e^{-\lambda t}\,\frac{(\lambda t)^x}{x!}, \quad x = 0, 1, 2, \cdots$$

and

$$P_\lambda(Y = y) = e^{-\lambda}\frac{\lambda^y}{y!}, \quad y = 0, 1, 2, \cdots$$

where Y, as above, is the demand during a lead period.

Suppose g is of the form

$$g(\lambda) = \begin{cases} \alpha \cdot \dfrac{(\alpha\lambda)^{n-1}}{\Gamma(n)} e^{-\lambda\alpha} & \text{for } \lambda > 0 \\ 0 & \text{for } \lambda \leq 0 \end{cases} \tag{16}$$

the so-called Gamma density with $\alpha > 0$ and $n > 0$. The general theory of Bayes solutions (Ref. 3) characterizes the Bayes solution for a given *a priori* density g as follows: if $X = x$ is observed, then the optimal choice $s(x)$ is that value of s which minimizes

$$\int_0^\infty \left\{hE_\lambda(s - Y) + \left(\frac{p}{T} + h\right)E_\lambda(Y - s)^+\right\} g_x(\lambda)\,d\lambda =$$

$$h\int_0^\infty E_\lambda(s - Y)\,g_x(\lambda)\,d\lambda$$

$$+ \left(\frac{p}{T} + h\right)\int_0^\infty E_\lambda(Y - s)^+ g_x(\lambda)\,d\lambda \tag{17}$$

the so-called *a posteriori risk*, where

$$g_x(\lambda) = \frac{\left(\dfrac{e^{-\lambda t}(\lambda t)^x}{x!}\right)\alpha\dfrac{(\alpha\lambda)^{n-1}}{\Gamma(n)}e^{-\lambda\alpha}}{\displaystyle\int_0^\infty \left(\frac{e^{-\lambda t}(\lambda t)^x}{x!}\right)\alpha\frac{(\alpha\lambda)^{n-1}}{\Gamma(n)}e^{-\lambda\alpha}\,d\lambda}$$

which is called the *a posteriori* density of λ given x. A routine computation reveals that $g_x(\lambda)$ is a Gamma density with parameters $\alpha + t$ and $n + x$. Thus, the first integral in Eq. (17) can be written

$$\int_0^\infty \left(\sum_{y=0}^\infty (s - y)e^{-\lambda}\frac{\lambda^y}{y!}\right)(\alpha + t)\frac{[(\alpha + t)\lambda]^{n+x-1}}{\Gamma(n + x)}e^{-\lambda(\alpha+t)}\,d\lambda = \sum_{y=0}^\infty \frac{(s - y)}{y!\,\Gamma(n + x)}\int_0^\infty e^{-\lambda(\alpha+t+1)}(\alpha + t)^{n+x}\lambda^{n+x+y-1}d\lambda =$$

$$\sum_{y=0}^\infty (s - y)\cdot\binom{n + x + y - 1}{y}\left(\frac{1}{1 + \alpha + t}\right)^y\left(\frac{\alpha + t}{1 + \alpha + t}\right)^{n+x}$$

This is the expectation of $s - Y$ when Y has distribution

$$P(Y = y) = \binom{n + x + y - 1}{y}\left(\frac{1}{1 + \alpha + t}\right)^y\left(\frac{\alpha + t}{1 + \alpha + t}\right)^{n+x}, \quad y = 0, 1, \cdots \tag{18}$$

which is the negative binomial distribution with parameters $n + x$, $(1 + \alpha + t)^{-1}$. Similarly, the second integral in Eq. (17) is the expectation of $(Y - s)^+$ for the same distribution of Y, which we call the *a posteriori* distribution of Y. Just as in Eq. (12), therefore, the minimizing s is the qth quantile of the distribution, i.e., the Bayes solution uses for $s(x)$ the smallest s so that

$$\sum_{y=0}^{s}\binom{n + x + y - 1}{y}\left(\frac{1}{1 + \alpha + t}\right)^y\left(\frac{\alpha + t}{1 + \alpha + t}\right)^{n+x} > q = \frac{\dfrac{p}{T}}{\dfrac{p}{T} + h}$$

In the general case where orders are of multiple sizes, the Bayes solutions are readily obtained for *a priori* densities on λ which are products of individual Gamma densities on $\lambda_1, \lambda_2, \cdots, \lambda_m$ (assuming $\lambda_{m+1} = \lambda_{m+2} = \cdots = 0$). These *a priori* densities result in an *a posteriori* distribution of Y which can be obtained as the distribution of the sum

$$Y_1 + 2Y_2 + \cdots + mY_m$$

where $Y_1, \cdots, Y_m$ are independent (representing the number of orders of sizes $1, \cdots, m$) and each $Y_k\ (k = 1, \cdots, m)$ has negative binomial distribution with parameters $n_k + x_k$, $(1 + \alpha + t)^{-1}$, (n_k, α being the parameters of the *a priori* Gamma density of λ_k). The distribution of Y can therefore be calculated by applying the following recursion formula for $k = 1, \cdots, m - 1$ successively:

$$P(Y_1 + 2Y_2 + \cdots + (k+1)Y_{k+1} = y) =$$

$$\sum_{j=0}^{[y/(k+1)]} P(Y_{k+1} = j)\, P(Y_1 + \cdots + kY_k = y - (k+1)j)$$

where $[y/(k+1)]$ denotes the largest integer $\leqq y/(k+1)$. The probabilities $P(Y_{k+1} = j)$ come from the negative binomial distribution, Eq. (18), with parameters

$$n_{k+1} + x_{k+1}, \quad (1 + \alpha + t)^{-1}$$

As before, the Bayes solution chooses s as the qth quantile of the *a posteriori* distribution of Y.

This *a posteriori* distribution of Y can be expressed in another way which is useful for computation and for intuitive appreciation of how the choice of the parameters of the *a priori* distribution affects the determination of s. It is helpful to start with the case where all the n_k's are positive integers. Consider three consecutive time periods of lengths $\alpha, t, 1$ (measured in units of one lead period, as above). Reasoning that any order of size k arriving during the total period of length $\alpha + t + 1$ has probability $1/(\alpha + t + 1)$ of arriving during the period of length one, we have a game of "heads and tails" where each order of size k (k fixed) is scored as "heads" if it lands in the time period of length one, and "tails" if it lands in the preceding time period of length $\alpha + t$. Our situation is this. For orders of size k, suppose we observe n_k during the time period α, then x_k during time period t. We then have a total of $n_k + x_k$ tails, but we do not know the number of heads (orders of size k during time period 1). However, since Prob. (heads) $= 1/(\alpha + t + 1)$, the probability of getting m heads before $n_k + x_k$ tails can be calculated for $m = 0, 1, 2, \cdots$ and leads to the negative binomial distribution with parameters $n_k + x_k$, $(1 + \alpha + t)^{-1}$. This is precisely the distribution we found above for Y_k, the number of orders of size k in the construction of the *a posteriori* distribution of Y! Moreover, this interpretation of the negative binomial distribution for Y_k makes clear the fact that Y_k can be constructed as the sum of $n_k + x_k$ independent and identically distributed variables: the number of heads before the first tail, the number between the first and second tails, $\cdots$, the number between the $(n_k + x_k - 1)$-th and $(n_k + x_k)$-th. Each of these has the geometric distribution

$$P(j) = \left(\frac{1}{1 + \alpha + t}\right)^j \left(\frac{\alpha + t}{1 + \alpha + t}\right), \quad j = 0, 1, 2, \cdots$$

We have thus obtained a heuristic interpretation of the distribution of the Y_k's (and, through them, an interpretation of the *a posteriori* distribution of Y). One specifies n_k orders of size k as occurring during an "*a priori* time period" of length α, one observes x_k orders of size k during a time period of length t, and from each of these $n_k + x_k$ orders of size k, one infers a geometrically distributed number of orders of size k during the lead period (on length one). (The mean of this geometric distribution is $(\alpha + t)^{-1}$, incidentally, so that the expected value of Y_k is $(n_k + x_k)/(\alpha + t)$, which is the average number of orders of size k per unit time during the combined α and t time periods).

The point of the heuristic discussion above is not to justify the Bayes solutions, whose minimization of average regret is justification enough. Apart from making the modus operandi of the Bayes solution appear plausible, the discussion is important in that it yields two important results:

(1) The effect of α and the n_k's is as though one observed n_k orders of size k during an "*a priori* time period" of length α. Since the n_k's need only be positive, not necessarily integers, this interpretation applies in the general case "by interpolation."

(2) The effect of α and the n_k's diminishes to zero as observations are continued over a time period $t \to \infty$. (For example, the expected value of Y_k, $(n_k + x_k)/(\alpha + t)$, is asymptotic to x_k/t.)

A practical difficulty in working with the Bayes solutions obtained above is that one must specify not only α, but also the values of n_k, i.e., the number of orders of each size k during the "*a priori* time period" α. How large should k be allowed to be in this specification? If there are many possible values of k, then what recipes can be used to prescribe the n_k's? These are difficult questions, which suggest that a considerable practical advantage can be realized by approximating the Bayes procedures for setting s by a family of procedures involving a small, fixed number of parameters. To this end, it is sufficient to find a good approximation to the *a priori* distribution of Y, which comes from convolutions of negative binomial distributions. Approximate Bayes solutions are then obtainable by "updating" the *a posteriori* distribution of Y each time an actual order occurs during the time period t (in the manner described above). Investigation of a variety of cases indicated that satisfactory approximations were obtainable by replacing the specified *a priori* distribution of Y by a gamma distribution having the same mean and variance. (Actually, since the gamma distribution is continuous, one "discretizes" it by taking the probabilities of unit-length intervals). This was checked out by comparing the regret functions of the approximate Bayes solutions with those of the corresponding Bayes solutions.

We have now a class of (approximate Bayes) procedures for setting s for a CSF. Each member of this class has four parameters associated with it:

q = the quantile of the *a posteriori* distribution of Y that is chosen for s, the cost-criticality quantile

α = the length of the "*a priori* time period"

G = mean demand specified for *a priori* period

σ^2 = variance of demand specified for *a priori* period

It is convenient to use the parameters $G_1 = \sigma^2/G$ rather than σ^2. The question that remains is how to choose the four parameters q, α, G, G_1 for individual items and CSFs. The relation of Eq. (14) showed that

$$q = \frac{1}{1 + \frac{hT}{p}} \tag{19}$$

is the optimal choice.

After setting the cost parameters and determining T approximately by Eq. (4), the choice of q by Eq. (19) is determined by the choice of p. The standard approach to inventory models (Ref. 2) regards p as a "penalty cost," not actually paid, but equivalent in dollars to the negative consequences of each unfilled demand. For some items it may not be difficult to estimate such a figure. For others, another approach may be easier, namely, to choose q itself directly. The latter can be accomplished by computing, for a range of specified q-values, both the expected shortage and the inventory level or, equivalently, the expected inventory cost. Thus the tradeoff between the two can be judged directly. An illustration of this approach is given in Appendix A.

With either of the methods for determining q, there is the possibility of failing to choose the truly optimal q. Both methods involve T, the mean length of the order cycle, whose determination is based in part on estimates of CSF mean demand rates. The error of these estimates is therefore reflected in the choice of q. The effect of this error on the regret was studied and typical results are discussed in Appendix B. It was found that the increased regret from this source should be relatively small. In fact, even a mis-estimate by a factor of 2 does not cause very large additional regret. By the same token, the tradeoff between expected shortage and inventory cost does not have to be determined precisely in order to achieve near-optimum results.

C. Evaluation of Procedures for Determining s

The choices of α, G, and G_1 are potentially quite variable. The terms G and G_1 can be regarded as postulated values (perhaps "engineering estimates") of the mean demand per lead time and variance/mean ratio. (If the demand is compound Poisson, then G_1 is the average size of the order to which a randomly selected unit belongs.) The parameter α measures how much weight is placed upon these postulated values. In other words, a large value of α tends to reduce the regret in those instances where the true mean demand and variance/mean ratio are close to G and G_1, respectively, but increases the regret if the true values are substantially different.

To investigate the consequences of various choices of the parameters α, G, and G_1, a computer program was developed to compute to within any specified accuracy a representative set of values of the regret function for a given set of parameter values and given t. Extensive investigations were made in the cases $t = 2, 4,$ and 6 because this range (in the unit of time equal to one lead period) is likely to be observed for current DSN items before their first reorder is necessary. The regret function computations (which also yielded computations of expected shortages and average inventory levels) were carried out for (true) mean demands ranging from 0.5 items per lead period (per CSF) to 30.5 items per lead period. This demand was constituted of orders of sizes 1 and 5 or 1, 3, and 10, with varying frequencies of order size to the overall demand. Thus cases like "infrequent order sizes greater than one," "frequent large order sizes," "mainly moderate order sizes" were investigated systematically, with q taking the values 0.85, 0.90, 0.95, 0.97, 0.99.

The following guidelines were developed from these investigations:

(1) If mean demand can be estimated with high confidence of no more than a factor of 3 error (in either direction), then the range $\alpha = 0.4$ to 0.6 is most effective in minimizing regret.

(2) The value of G should be chosen 25–50% higher than the estimated mean demand.

(3) The value of G_1 should be chosen 40–80% higher than the estimated variance/mean ratio for demand during lead time.

These seemingly conservative choices of G and G_1 are preferable because of the nature of the regret functions for the procedures under consideration. The choice of G results from the fact that it is relatively difficult to minimize regret for higher demand rates. The choice of G_1 on the high side also helps keep the regret small for higher demand rates and, in addition, is particularly useful in protecting against sporadic demand for larger than normal order sizes. Where this sporadic demand is considered a greater possibility, still larger values of G_1 are called for.

Besides developing the guidelines for choosing specific procedures for setting CSF minimum levels, the computational investigation was aimed at evaluating the performance of this type of procedure. Specifically, it was desired to see whether the regret was fairly uniformly controlled for large ranges of demand distribution parameters. It became readily apparent that the regret tended to rise directly with the demand level. In other words, the regret tends to be a fairly constant percentage of the minimum possible overall cost (Eq. 2). (Recall that this minimum possible overall cost is attainable only if the true demand distribution is known.) In Fig. 2, the percentage regret is plotted as a function of mean demand during lead time. A separate curve is plotted for each of four cases (labeled "Frequency of Order Sizes"). It is assumed that orders are of two sizes, one unit and five units, and the cases specify the frequency of each order size.

Note that the percentage regret in all four cases is about 15% for mean demand in the range 8.5 to 30.5, and is somewhat larger, about 20–25%, for smaller demand levels. (Of course, the regret in absolute terms is still substantially smaller for low demand than for higher demand.) This mild increase in percentage of regret at the extreme lower end of the demand range is typical, and results largely from the conservative effect of the *a priori* estimates G and G_1, which causes the minimum levels to provide reasonable protection against shortages even if the observed demand is quite small. Figure 3 shows the results of a more elaborate investigation of the performance of the same procedure ($\alpha = 0.4$, $G = 15$, $G_1 = 7$, $q = 0.90$).

The eight cases shown in Fig. 3 represent variations of the frequency of order sizes for orders of sizes 1, 3, and 10.

Note that the regret is in the 15–20% range, approximately the same as in Fig. 2. Both figures give the results for $t = 4$ lead periods, which is on the order of one year's experience with observed demand. The range of 15–20% regret is the best that can be accomplished by any of the procedures studied uniformly over a broad range of possible mean demand levels and variations in frequency of order sizes. It is appropriate to consider this regret as the (essentially unavoidable) cost of the degree of uncertainty regarding the true demand distribution that remains after one year's observation. Of course, considerable improvement in reducing overall costs is possible in the long term if the system is implemented in a consistent fashion and the demand distributions are fairly stable. As an illustration of the improved percentage of regret typical after 1½ year's observation ($t = 6$), we have Fig. 4.

Here the pattern is similar to Fig. 2, but the regret is mainly in the 10–15% range rather than in the 15–20% range. The computations of regret become more time-consuming for large t (the time increasing with t^2), but routine mathematical arguments show that, as would be expected, the percentage regret over any fixed demand range approaches zero uniformly as t approaches infinity.

References

1. *The Economic Order Quantity Principle and Applications, A GSA Handbook*, FMPR 101-27.1, General Services Administration, Washington, D.C., May 1966.

2. Wagner, H. M., *Principles of Operations Research*, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1969, Chapter 19.

3. Wald, A., *Statistical Decision Functions*, Wiley and Sons, New York, New York, 1950.

Table 1. Excerpts from a typical allocation table

Q			$k = 6$			CSF annual demand rates 7:8:9:10:11:12		
	n_1	n_2	n_3	n_4	n_5	n_6	Expected time	Expected residual
6	1	1	1	1	1	1	.018	5.00
7	1	1	1	1	1	2	.021	5.79
8	1	1	1	1	2	2	.026	6.53
9	1	1	1	2	2	2	.032	7.18
10	1	1	2	2	2	2	.040	7.71
⋮								
40	6	6	6	7	7	8	.396	17.41
⋮								
105	14	15	17	18	20	21	1.32	29.54
⋮								
117	16	17	19	20	22	23	1.50	31.35
⋮								
142	19	21	23	25	26	28	1.88	34.82

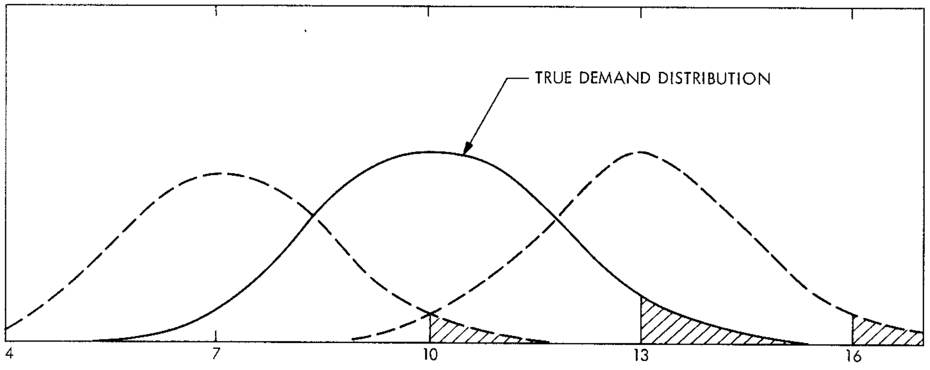
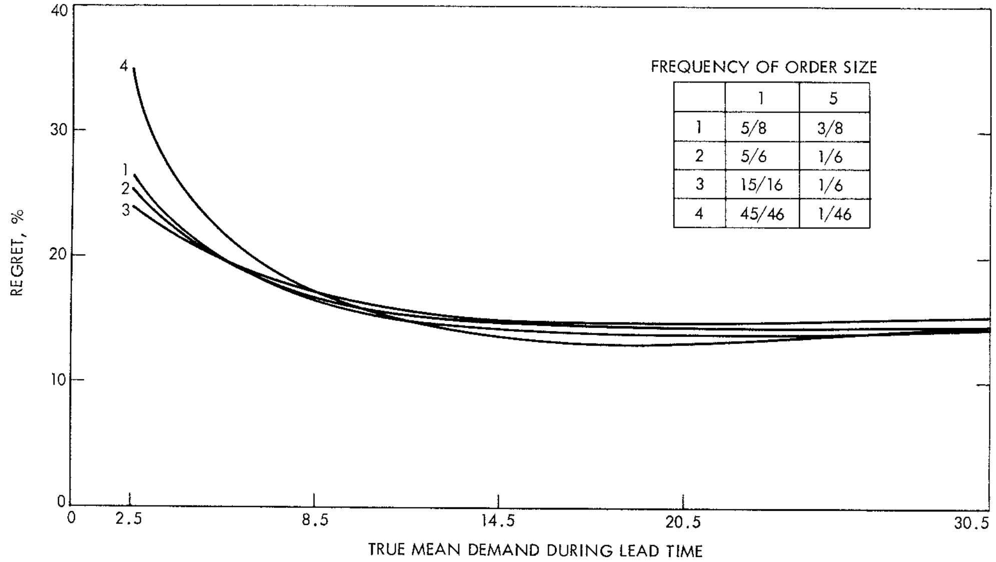


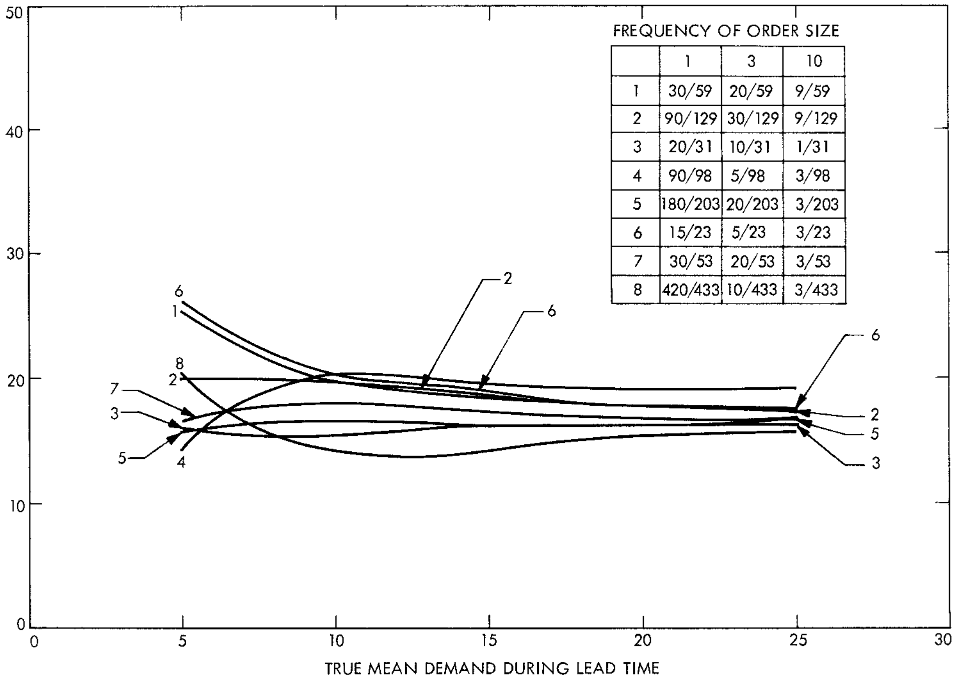
Fig. 1. Setting minimum levels at CSF



FREQUENCY OF ORDER SIZE

	1	5
1	5/8	3/8
2	5/6	1/6
3	15/16	1/6
4	45/46	1/46

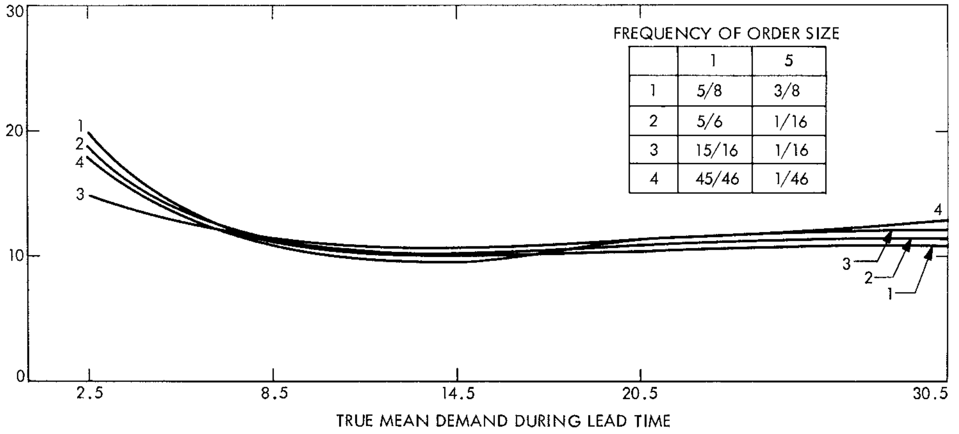
Fig. 2. Regret as a percentage of average cost for a range of demand distributions ($t = 4$, $\alpha = 0.4$, $G = 15$, $G_1 = 7$, $q = 0.90$)—order sizes 1, 5



FREQUENCY OF ORDER SIZE

	1	3	10
1	30/59	20/59	9/59
2	90/129	30/129	9/129
3	20/31	10/31	1/31
4	90/98	5/98	3/98
5	180/203	20/203	3/203
6	15/23	5/23	3/23
7	30/53	20/53	3/53
8	420/433	10/433	3/433

Fig. 3. Regret as a percentage of average cost for a range of demand distributions ($t = 4$, $\alpha = 0.4$, $G = 15$, $G_1 = 7$, $q = 0.90$)—order sizes 1, 3, 10



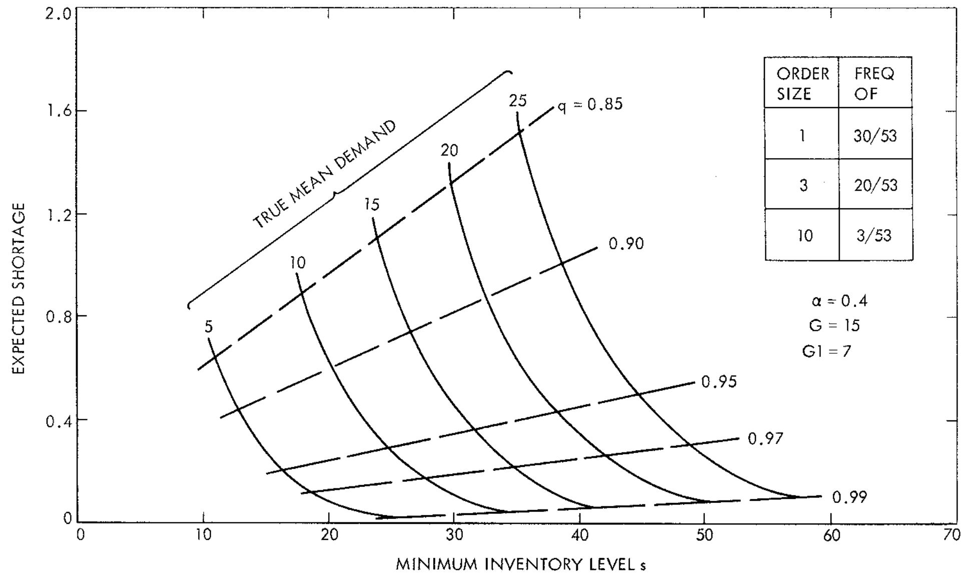
FREQUENCY OF ORDER SIZE

	1	5
1	5/8	3/8
2	5/6	1/16
3	15/16	1/16
4	45/46	1/46

Fig. 4. Regret as a percentage of average cost for a range of demand distributions ($t = 6$, $\alpha = 0.6$, $G = 10$, $G_1 = 7$, $q = 0.90$)—order sizes 1, 5

Appendix A
Tradeoff of Expected Shortage vs Inventory Level

Each curve in Fig. A-1 corresponds to a different (true) mean demand level. The five points distinguished on each curve correspond (moving from left to right) to $q = 0.85, 0.90, 0.95, 0.97$, and 0.99. Intermediate points can be attained by choosing intermediate values of q. Moving along a fixed curve from left to right, the expected shortage decreases (as q is increasing) while the average value of s, the minimum stockage level, increases. Increments of small s when multiplied by h give a good approximation to increments in average inventory cost per year. Thus the curves indicate how much decrease in expected shortage is "bought" for arbitrary levels of increased inventory levels. The "flattening out" of the curves shows that in the range of 0.97 to 0.99 a relatively small rate of decrease in expected shortage is obtained by increasing the inventory level.



ORDER SIZE	FREQ OF
1	30/53
3	20/53
10	3/53

Fig. A-1. Expected shortage vs inventory level and demand

Appendix B
Effect of Mis-estimating q

Figure B-1 shows the overall average cost (Eq. 2) as a function of the mean demand per lead period. The order sizes are 1, 3, and 10, with fixed frequency of order sizes. The bottom curve plots the minimum possible cost (attainable only if the demand distribution is known). The second curve from the bottom shows the average cost incurred when the correct choice $q = 0.95$ is made. (Thus the distance between the two curves is the regret.) The two upper curves represent the average cost incurred when an erroneous q is used (e.g., through mis-estimating T or setting p poorly).

The value $q = 0.97$ corresponds to an underestimate of T by a factor of 3/5, while the value of $q = 0.90$ corresponds to an overestimate by a factor of 2. Even these levels of error produce an increment to average cost substantially smaller than the regret.

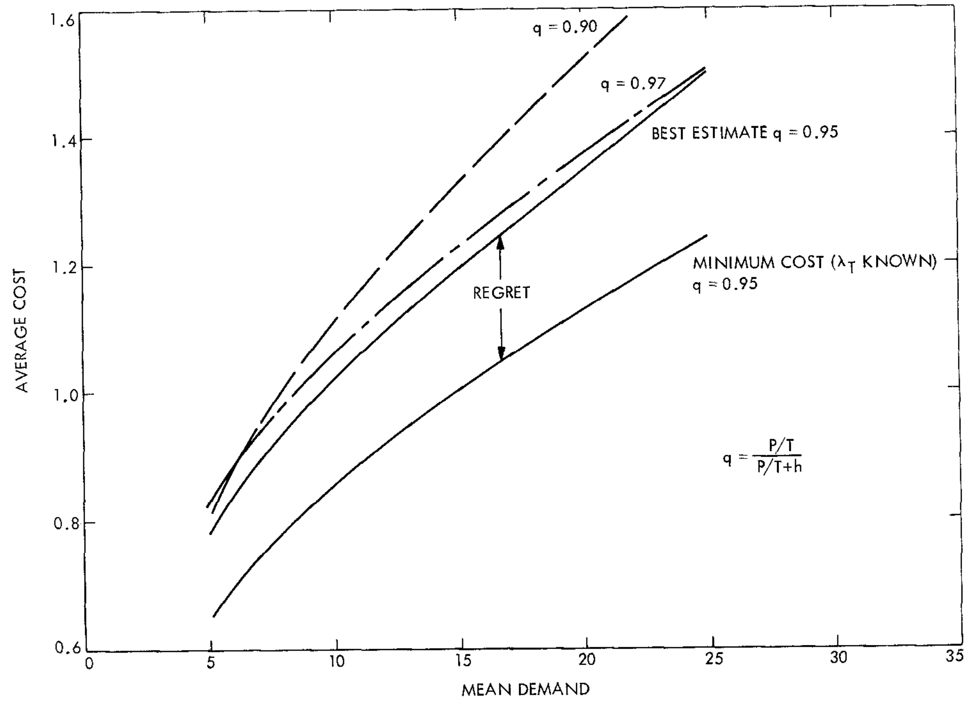


Fig. B-1. Effect of mis-estimating *q*